Data Science Self-Assessment

Galvanize Inc.

Contents

How to Use This Document

This document is designed to give you an idea of the baseline of Python, SQL and probability/statistics knowledge required to apply for the Data Science Immersive program. If understanding any of the scripts included in this PDF is challenging, we encourage you to take the time to study Python and/or SQL and/or probability/statistics before beginning the application process. For a list of free Python, SQL and probability/statistics resources, please refer to the DSI Study Resources PDF.

This document starts with some simple python statements which you should be able to evaluate without actually executing. We then proceed to more advanced challenges that will require a solid understanding of strings, lists, sets, dictionaries, file I/O, and functions. We then continue the self assessment with a variety of SQL statements you should be comfortable with. We end the document with probability/statistics exercises that cover counting (permutations, combinations), probability (conditional probability, Bayes' Theorem), probability distribution for discrete and continuous random variables, descriptive and inferential statistics as well as basic linear regression.

Spot the Differences

Without running the scripts, can you tell what the output will be? If you have some Python or programming background, this section should take very little time.

For Loops

```
# Script 1
list_num = [1,2,3]
for num in list_num:
    total = 0
    total += num
    print total
```

```
# Script 2
list_num = [1,2,3]
total = 0
for num in list_num:
    total += num
    print total
```

```
# Script 3
list_num = [1,2,3]
total = 0
for num in list_num:
    total += num
print total
```

For Loops in Functions

```
# Script 1
def my_function1(my_list):
    output = []
    for item in my_list:
        output.append(item)
        return item

print my_function1([’cat’, ’bad’, ’dad’])
```

```
# Script 2
def my_function2(my_list):
    output = []
    for item in my_list:
        output.append(item)
        return output

print my_function2([’cat’, ’bad’, ’dad’])
```

```
# Script 3
def my_function3(my_list):
    output = []
    for item in my_list:
        output.append(item)
    return item

print my_function3([’cat’, ’bad’, ’dad’])
```

```
# Script 4
def my_function4(my_list):
    for item in my_list:
        output = []
        output.append(item)
    return output

print my_function4([’cat’, ’bad’, ’dad’])
```

—

```
# Script 5
def my_function5(my_list):
    output = []
    for item in my_list:
        output.append(item)
    return output

print my_function5([’cat’, ’bad’, ’dad’])
print my_function5([’cat’, ’bad’, ’dad’])
```

```
# Script 6
output = []
def my_function6(my_list):
    for item in my_list:
        output.append(item)
    return output

print my_function6([’cat’, ’bad’, ’dad’])
print my_function6([’cat’, ’bad’, ’dad’])
```

—

Make a function

Functions, blocks of reusable code, keep your code modular, well organized and easily maintainable. You should try to keep your code organized in functions. Take a look at each of the following snippets of code and organize them into functions.

1. We want a function that takes a list of numbers and returns that list where 10 was added to each number.

```
list_num = [1,2,3]
list_add_10 = []
for num in list_num:
    list_add_10.append(num + 10)
print list_add_10
```

2. We want a function that takes in a list of strings and returns the list with the length of the words.

```
list_words = [’great’, ’job’, ’so’, ’far’]
list_length_words = []
for word in list_words:
    list_length_words.append(len(word))
print list_length_words
```

More Advanced Python Challenges

Practice, practice, practice: we encourage you to work through these challenges.

Challenge 1

Write a function that looks at the number of times given letters appear in a document. The output should be in a dictionary.

~~~python
def letter_counter(path_to_file, letters_to_count):
    ’’’ Returns the number of times specified letters appear in a file

    Parameters
    -----------
    path_to_file: str
        Relative or absolute path to file of interest
    letters_to_count: str
        String containing the letters to count in the text

    Returns
    --------
    letter_dict: dict
        - key: letter
        - value: the count of that letter in the file
    The counting is case insensitive

    Example
    --------
    ```file.txt
    This is the file of interest. Count my vowels!
    ```
    >>> letter_counter(’file.txt’, ’aeiou’)
    {’i’: 4, ’e’:4, ’o’:2, ’u’:1}
    ’’’
    pass
~~~

Challenge 2

Write a function that removes one occurrence of a given item from a list. Do not use methods `.pop()` or `.remove()`! If the item is not present in the list, output should be ‘The item is not in the list’.

```python
def remove_item(list_items, item_to_remove):
    ''' Remove first occurrence of item from list

    Parameters
    ----------_
    list_items: list
    item_to_remove: object
        The object to be removed form list_items

    Returns
    --------
    - if the item is in the list: list
        list with first occurrence of item removed
    - if the item is not in the list: str
        'The item is not in the list'

    Example
    --------
    >>>list_items = [1,3,7,8,0]
    >>>remove_item(list_items, 7)
    [1,3,8,0]
    '''
    pass
```

Challenge 3

The simple substitution cipher basically consists of substituting every plaintext character for a different ciphertext character. The following is an example of one possible cipher from http://practicalcryptography.com/ciphers/simple-substitution-cipher/:

- Plain alphabet : abcdefghijklmnopqrstuvwxyz
- cipher alphabet: phqgiumeaylnofdxjkrcvstzwb

```
def cipher(text, cipher_alphabet, option=’encipher’):
    ’’’ Run text through a particular cipher alphabet

    Parameters
    -----------
    text: str
        Either the plain text to encipher, or the cipher text to decrypt
    cipher_alphabet: dict
        Dictionary specifying {’original_letter’: ’cipher_letter’}
    option: str (default ’encipher’)
        ’encipher’ (accept plain text and output cipher text)
        ’decipher’ (accept cipher text and output plain text)

    Returns
    --------
    cipher text by default,
    plain text if option is set to decipher

    >>> d = dict(zip(’abcdefghijklmnopqrstuvwxyz’,
                     ’phqgiumeaylnofdxjkrcvstzwb’))
    >>> cipher(’defend the east wall of the castle’,
               d)
    ’giuifg cei iprc tpnn du cei qprcni’
    >>> cipher(’giuifg cei iprc tpnn du cei qprcni’,
               d,
               option=’decopher’)
    ’defend the east wall of the castle’
    ’’’
    pass
```

Challenge 4

Implement a function that counts the number of isograms in a list of strings.

- An isogram is a word that has no repeating letters, consecutive or non-consecutive.
- Assume the empty string is an isogram and that the function should be case insensitive.

```python
def count_isograms(list_of_words):
    ''' Count the number of strings without repeating characters in a list

    Parameters
    -----------
    list_of_words: list of strings

    Returns
    -------
    count of isograms (as integer)

    >>>count_isograms(['conduct',  letter', 'contract', 'hours', 'interview'])
    1
    '''
    pass
```

Challenge 5

Write a function that returns a list of matching items. Items are defined by a tuple with a letter and a number and we consider item 1 to match item 2 if:

1. Both their letters are vowels (aeiou), or both are consonnants and,
2. The sum of their numbers is a multiple of 3

(1,2) contains the same information as (2,1), the output list should only contain one of them.

```python
def matching_pairs(data_list):
    '''
    Parameters
    ----------
    data_list: as list of tuples (letter, number)

    Returns
    -------
    A list of the matching pair referenced by their index (index_A, index_B).
    Each pair should appear only once. (A,B) is the same as (B,A)

    >>> data = [('a', 4), ('b', 5), ('c', 1), ('d', 3), ('e', 2), ('f',6)]
    >>> matching_pairs(data)
    [(0,4), (1,2), (3,4)]
    '''
    pass
```

Getting Ready for the SQL Assessment

You should be able to write the SQL queries that use `SELECT`, `FROM`, `WHERE`, `CASE` clauses, aggregates, and `JOIN`s . To check your work, you can run your queries on w3school's site (http://bit.ly/1foSkgu)

Our Data

We will be querying the following tables.

Table 1: `flags`

name	country	w_prop	l_prop	adoption_date
"Tricolour"	"France"	2	3	1830
"Union Jack"	"United Kingdom"	1	2	1801
"The Star-Strangled Banner"	"USA"	10	19	1960
"Hinomaru"	"Japan"	2	3	1999
"NA"	"Brazil"	7	10	1992
"Jalur Gemilang"	"Malaysia"	1	2	1963

where `w_prop` is the width proportion and `l_prop` is the length proportion

Table 2: `countries`

country	capital	contient
"France"	"Paris"	"Europe"
"Malaysia"	"Kuala Lumpur"	"Asia"
"Brazil"	"Brasilia"	"South America"
"United Kingdom"	"London"	"Europe"
"Japan"	"Tokyo"	"Asia"
"USA"	"Washington DC"	"North America"
"Germany"	"Berlin"	"Europe"

Simple Queries on a Single Table

1. Use the `WHERE` clause to show the countries with a flag ratio of 2:3 (i.e. `w_prop` = 2 and `l_prop` = 3).
2. Use `IN` to check if an item is in a list and show the countries on a continent that is either Europe or North America.
3. Use `BETWEEN xxx AND xxx` to show names of flags and countries that have width proportion higher than 1 but lower than 8.
4. Use `LIKE 'X%'` to show countries that have an name that starts with 'U'.
5. Use `CASE` to show countries, their capital and a column to indicate whether the continent is 'Eurasia' (i.e. Europe or Asia) or 'Americas' (North or South America). Add a filter to select countries with capitals that are at least 7 character long.

Build Queries with Aggregates

Aggregates include commands such as `DISTINCT`, `COUNT`, `SUM`, `GROUP BY`, `HAVING`, and `ORDER BY`. Try using these commands on the following questions!

1. Use `DISTINCT` to list the continents in the countries table - each continent should appear only once.
2. Use `COUNT` to see how many countries are in Europe.
3. Use `GROUP BY` to count how many countries are in each continent, with continents alphabetically ordered (hint: use `ORDER BY`).
4. Use `HAVING` to determine which continents are represented at least twice in the countries table.

Build Complex Queries on Multiple Tables

1. Use `JOIN` to display the capital, the country, and the flag name.
2. Use `JOIN` and `WHERE` to display the continents associated to the flags in the flags table when the flag has a name (i.e. not 'NA').
3. Use `JOIN` and `HAVING` to display continents that have at least 2 countries represented as well as the average adoption date of the flag (as `avg_date`).

Practice some probability and statistics

Here is a small selection of exercises to make sure you know how to apply your knowledge in statistics, probability and simple regression. If you want to practice some more, or to practice on exercises with a solution, checkout the links in each section. They come from the recommended resources (Khan Academy, Udacity and the probability review).

Table of content - Counting: permutations, combinations

- Probability: Probability of an event, Probability of 2 or more events (Conditional probability, Independent and dependent events, Mutually exclusive events, Bayes' Theorem)
- Probability distribution (Binomial, Geometric and Poisson distributions for discrete random variables, Uniform, Normal and Exponential distributions for continuous random variables)
- Descriptive Statistics: mean, variance, standard deviation, range, IQR
- Inferential Statistics: confidence intervals, hypothesis testing, inference for proportions and means
- Linear regression: model performance, interpretation of coefficients, underfitting/overfitting

NB: some exercises are labeled as *extra credit*, and as such are not mandatory.

Counting: permutations, combinations

1. Permutations

1. How many ways can you arrange the numbers 1, 2, 3, 4 and 5?
2. How many ways can you arrange 1, 1, 2, 3, 4?
3. How many ways can you arrange two 3s and three 5s?

some links: http://bit.ly/2iGgrir, http://bit.ly/2jXtFIt

2. Combinations

1. How many different poker hands (5 cards) can you have? A deck holds 52 cards.
2. There are five flavors of ice cream: Stracciatella, Mint chocolate chip, Cookies and Cream, Butter Pecan, Pistachio and Pralines and cream. How many three scoop ice-creams can you make if all the scoops must be different flavors?

For extra credit: what happens if you can take several scoops of the same flavor?

some links: http://bit.ly/2iNIXSF, http://bit.ly/2jXlDiI

Probability

1. Probability of an event

1. In a deck of cards (52 cards), what's the probability of picking a queen? a heart? Of picking a card that's not a queen nor a heart?
2. If I do not replace the cards, what is the probability of picking 2 kings? 4 diamonds? How do these probabilities evolve if I replace the cards after each draw?

some links: http://bit.ly/2iNCwyS, http://bit.ly/OtSNH2, http://bit.ly/2j7R4qF

2. Probability of 2 or more events

Conditional probability

1. What is the probability that the total of two dice is less than four, knowing that the first die is a 2?
2. 90% of candidates to a Web developer position can code both in Javascript and HTML. 70% of these candidates can code in Javascript and 50% can code in HTML. What is the probability that a candidate can code in HTML knowing that he can code in Javascript?

some links: http://bit.ly/2iGktHi

Independent and dependent events

1. Number of kids dressed as pumpkins vs ghost that obtained a certain number of pieces of candy on Halloween night for a kid dressed up as a pumpkin

Number of candy	less than 10	from 10 to 20	from 20 to 30	more than 30
Pumpkins	5	10	60	25
Ghosts	15	40	80	15

- What is the probability that a kid dressed as a pumpkin gets 20 or more pieces of candy? How about if he dresses as a ghost?
- What is the probability that a kid obtains less than 10 pieces of candy?
- What is the probability that two siblings, one dressed as a ghost and one dressed as a pumpkin, each receive 20 to 30 pieces of candy?

2. You toss a fair die twice. What is the probability of getting less than 3 on the first toss and an even number on the second?

some links: http://bit.ly/2jmalpl

Mutually exclusive events

Let's consider a population from which we draw a sample of 40 individuals. You know that the probability of not obtaining anyone wearing glasses in the sample in 26%. The probability of having only one individual wearing glasses is 32%. What is the probability of

(a) obtaining not more than one individual wearing glasses in a sample?

(b) obtaining more than one individual wearing glasses in a sample?

some links: http://bit.ly/2jmjyxO

Bayes' Theorem

1. To detect a medical condition, patients are given two tests. 25% of the patients receive positive results on both tests and 42% of the patients receive positive results on the first test. What percent of those who have positive results on the first test passed also had positive result on the second test?

2. Extra Credit: A jar contains red and blue marbles. You draw two marbles one after the other without replacing the first marble in the jar. You know that:

- The probability of selecting a blue marble and then a red marble is 30%.
- The probability of selecting a red marble on the first draw is 50%.

You first draw a red marble. What is the probability of selecting a blue marble on the second draw?

some links: http://bit.ly/2jmjHRS

Probability distributions Problems

Common problems relying on discrete (Binomial, Geometric, Poisson) or continuous (Uniform, Normal, Exponential) probability distributions.

Here are some exercises with their solutions as video.

Binomial distribution

1. Fair coin: Imagine you were to flip a fair coin 10 times. What would be the probability of getting 5 heads?

2. Unfair coin: You have a coin with which you are 2 times more likely to get heads than tails. You flip the coin 100 times. What is the probability of getting 20 tails? What is the probability of getting at least one heads?

Geometric distributions

Suppose you have an unfair coin, with a 68 % chance of getting tails. What is the probability that the first head will be on the 3rd trial?

Poisson distribution

There are on average 20 taxis that drive past your office every 30 minutes. What is the probability that 30 taxis will drive by in 1 hour?

Exponential distribution

Let X, the number of years a computer works, be a random variable that follows an exponential distribution with a lambda of 3 years. You just bought a computer, what is the probability that the computer will work in 8 years?

extra credit: Let X be a random variable that now follows an exponential distribution with a half-life of 6 years. Find the parameter of the exponential distribution. What is the probability P(X > 10) and the conditional probability P(X > 20 | X > 10)?

Uniform distribution

Let the random variable X be the angle of a slice of pizza. The angle X has a uniform distribution on the interval [0,90]. What is the probability that your slice of pizza will have an angle between 30 and 40°?

extra credit: X is uniform on the interval [a,b], can you derive the expected value E(X)? the variance V(X)?

Normal distribution

1. Suppose X has a standard normal distribution. Compute P(X > 9), P(1< X < 3) and P(X > -3).
2. The weight in pounds of individuals in a population of interest has a normal distribution, with a mean of 150 and a standard deviation of 40. What is the expected range of values that describe the weight of 68% of the population (Hint: use the empirical rule)? Of the people who weigh more than 170 pounds, what percent weigh more than 200 pounds (Hint: this is conditional probability)?

Descriptive Statistics

3 Ms

Give the mean, median and mode of the following data:

(20,45, 68, 900, 57, 45, 33, 35, 45, 22)

Do you think the mean is a good summary statistic? Why or why not?

Variance, Range, IQR

Give the mean, the variance, the standard deviation, the range and the interquartile of range of the following data:

(20,45, 68, 900, 57, 45, 33, 35, 45, 22)

Discrete random variables

Give the expression of the mean and the variance for a discrete random variable X.

Continuous random variables

Give the expression of the mean and the variance for a continuous random variable X.

Inferential Statistics

1. Confidence intervals

1. We are polling to get approval rate of a president. Out of a population of 4 million, 6014 were surveyed and 3485 expressed their approval. Construct a 95% confidence interval for the approval rate of the president.
2. The weight of a random sample of 100 individuals from a population of interest was surveyed and yielded a sample average weight of 150 pounds and sample standard deviation of 20 pounds. Construct a 95% confidence interval for the average weight of the population.

2. General hypothesis testing

1. What is the definition of a significance level? of a p-value?
2. Would you use a one tailed or two tailed tests in the following cases:
   - investigating if women are paid less than men.
   - comparing the click-through rate of website when the 'subscribe' button is green vs when it is blue.
3. A man goes to trial. In a hypothesis testing framework, let's define the null hypothesis as *Not Guilty* and the alternative hypothesis as *Guilty*.
   - What type of error is made when the man is actually not guilty but verdict returned is guilty?
   - What type of error is made when the man is actually guilty but verdict returned is not guilty?

3. One sample hypothesis testing

1. We want the test the hypothesis that at least 68% of the Canadian population (aged 18+) went to the movies at least once in the past 12 months with a significance leval of 5%. We surveyed 4,000 respondents and found 3,012 did go at least once to the movies in the past 12 months. How would your conclusion compare if you only had 40 respondents, 30 of which went to the movies at least once in the past 12 months

some links: http://bit.ly/2jIM1h3

2. We want to test the hypothesis that the average weight in North America is at least 175 pounds. The mean of weights of the 100 individuals sampled is 178 pounds, with a sample standard deviation of 8 pounds. What are you conclusions?

some links: http://bit.ly/2jmht5d

3. We want to investigate the claim that on average, sea turtles lay 110 eggs in a nest. Volunteers have gone out and counted the number of eggs in 20 nest. What do you conclude?

- data: 101, 120, 154, 89, 97, 132, 126, 105, 94, 111, 98, 90, 88, 115, 99, 85, 131, 127, 116

some links: http://bit.ly/2j7KpN2

3. Two sample hypothesis testing

1. Is there a meaningful difference between the proportion of teenagers vs that of adults that go to the movies at least once per month?

- Data:
  - 1000 teenagers are surveyed, 780 answer positively.
  - 1000 adults are surveyed, 620 answer positively.

some links: http://bit.ly/2j7GUXg

2. Is there a meaningful difference between the average wingspan of bald eagles vs that of crowned eagles?

- data for bald eagles (in ft): [7.4, 7.7, 6.0, 6.7, 8.3, 6.5, 6.9, 7.7, 7.8, 7.3, 6.9, 6.5, 6.3, 4.8, 8.0, 6.8, 5.8, 6.9, 6.3, 6.3, 6.4, 5.1, 6.9, 7.6, 5.6, 6.5, 6.7, 7.8, 6.6, 6.9, 7.0, 6.4, 7.4, 6.0, 7.0, 5.3, 5.8, 6.4, 7.1, 5.5, 7.0, 6.7, 5.8, 6.1, 7.1, 7.9, 7.7, 6.2, 5.3, 6.4, 6.9, 5.9, 7.8, 5.6, 5.0, 5.5, 6.4, 7.1, 8.6, 9.3, 6.8, 7.6, 7.2, 7.1, 5.8, 5.9, 5.1, 6.6, 6.8, 5.7, 6.3, 7.3, 6.3, 7.2, 7.7, 6.0, 7.2, 5.9, 7.2, 7.0, 7.4, 6.5, 7.8, 5.9, 6.3, 6.3, 8.3, 5.9, 6.9, 7.8]
- data for crowned eagles (in ft): [5.3, 5.6, 5.8, 5.3, 5.6, 4.9, 5.7, 5.4, 5.8, 5.4, 6.0, 5.4, 5.1, 5.4, 5.2, 5.7, 4.8, 5.8, 5.7, 5.1, 5.3, 5.4, 5.7, 6.6, 5.0, 5.4, 5.3, 5.5, 5.2, 5.6, 5.2, 5.9, 5.7, 5.8, 5.5, 5.2, 4.0, 5.8, 5.2, 6.2, 5.4, 4.6, 5.3, 5.8, 6.3, 4.8, 5.6, 5.4, 5.2, 5.4, 5.1, 6.0, 6.1, 5.4, 5.4, 5.3, 5.0, 6.0, 5.0, 5.8, 5.1, 5.3, 4.8, 5.6, 5.7, 6.1, 5.0, 6.4, 5.1, 4.6, 5.3, 6.0, 4.8, 5.4, 4.3, 5.4, 5.1, 4.7, 6.0, 5.5, 5.4, 5.6, 5.2, 5.8, 5.3, 4.9, 5.3, 5.5, 5.7, 4.7, 6.0, 5.6, 4.9, 5.4, 4.3, 5.5, 4.9, 5.3, 5.6, 6.0]

some links: http://bit.ly/2jva7OY

Relationship between two quantitative variables

1. Dataset

x	0	1	2	3	5
y	1	2.1	3.2	4	6.1

   (a) Plot corresponding the scatter plot.

   (b) Find the least square regression line y = ax + b. Add it to your plot.

   (c) Estimate the value of y when x = 4.

Extra credit: Can you do these steps in Python?

2. Dataset

x	0	1	2	3	4	7	9	11	30
y	2.	4.9	8.	10.8	13.9	23.1	29.	35.	92.1

   (a) Find the least square regression line for the given data points.

   (b) Plot the given points and the regression line on the same graph.

3. We have the following (x,y) points: [(0, 42.0), (1, -101.0), (2, 21.0), (3, -38.0), (4, 5.0), (7, 20.0), (9, 293.0), (11, 266.0), (15, 625.0), (20, 1266.0), (25, 1757.0), (30, 2844.0)]

   (a) Plot the data.

   (b) How do you think a linear model would perform? How about a 100 degree polynomial model? How would you figure out which of these models was preferable?

   (c) How would you model the relationship between these features?

4. We have a dataset that gives the height and age of a sample of people. The range of age spans from 1 to 60 years. We decide to compute the correlation coefficient to model to understand the relationship between these features.

   (a) Do you expect the correlation coefficient to be positive or negative?

   (b) What are some of the limitation of this approach?

some links: http://bit.ly/2jXyDF6, http://bit.ly/2jqXuRp, http://bit.ly/2jxlCFA